### 

**Capstone Project – Data Wrangling**

**By Sumita Soundararajan**

***Foundations of Data Science Workshop by Springboard***

**Airbnb Recruiting - New User Bookings**

## ***Where will a new guest book their first travel experience?***



**Link to the original datasets ->**

<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data>

The original and cleaned datasets are too large to be attached to github.

**File descriptions**

The dataset we are researching is provided by Airbnb. The 5 datasets provided, contain a list of users along with their demographics, web session records, and some summary statistics.

1. train\_users.csv - the training set of users
2. test\_users.csv - the test set of users

## The training data consists of 213,451 rows with 16 columns while the test data has 62,096 rows with 15 columns. The values of 'country\_destination' are missing in the Test data and that is the value that is to be predicted.

* + id: user id
  + date\_account\_created: the date of account creation
  + timestamp\_first\_active: timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up
  + date\_first\_booking: date of first booking
  + gender
  + age
  + signup\_method
  + signup\_flow: the page a user came to signup up from
  + language: international language preference
  + affiliate\_channel: what kind of paid marketing
  + affiliate\_provider: where the marketing is e.g. google, craigslist, other
  + first\_affiliate\_tracked: what is the first marketing the user interacted with before signing up
  + signup\_app
  + first\_device\_type
  + first\_browser
  + country\_destination: this is the target variable to be predicted

1. sessions.csv - 10,567,737 rows of 6 variables with multiple rows per user-id. Has details on the web sessions log depicting the browsing behavior of users
   * user\_id: to be joined with the column 'id' in users table
   * action
   * action\_type
   * action\_detail
   * device\_type
   * secs\_elapsed

**Train.users.csv**

Totally 213,452 rows with 16 columns, with substantial missing values in column Age.

* In Age column, there are many values in thousands and some in single digits. Assume an age range of 15 to 100, and then assume the 4-digit years from 1924-1995 as birth years.
* Replace all ages outside the valid range to NA
* Replace all missing values to NA
* Calculate the 4-digit years as birth year by subtracting given year from 2016 (commencement of competition in Kaggle)

**Sessions.data.csv**

Totally 10,567,737 rows with 6 columns, with multiple rows for a user-id.

**User\_id**

1. Replace missing or blanks as “Not given”

**Action**

1. Replace “-unknown- “as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AC”) to differentiate between similarly named columns
4. Rename all Actions with Cumulative Percentage > 95% as MISC

**Action\_type**

1. Replace “-unknown- “as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AT”) to differentiate between similarly named columns
4. Rename all Action\_type with Cumulative Percentage > 98% as MISC

**Action\_detail**

1. Replace “-unknown- “as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AD”) to differentiate between similarly named columns
4. Rename all Action\_detail with Cumulative Percentage > 95% as MISC

**Device\_type**

1. Replace “-unknown- “as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_DT”) to differentiate between similarly named columns
4. Rename all Device\_type with Cumulative Percentage > 96% as MISC

**Secs\_elapsed**

1. Replace all NA or blanks to 0